The Indonesian Journal of Computer Science
www.ijcs.net
Volume 15, Issue 5, October 2026
https://doi.org/10.33022/ijcs.v15i5.5243

Toward a Computational Cognitive Index (CCI): Personalized EEG-Based Foundations for Multimodal Mental Workload Modeling

Jean Paul Kazungu [1], Jean Pierre Ntayagabiri [2*], M. Kokou Assogba[3], Jeremie Ndikumagenge[4]
jean-paul.kazungu@ub.edu.bi[1], jpntayaga2@gmail.com[2], kasogba1@avu.org[3], jeremie.ndikumagenge@ub.edu.bi[4]
[1]Doctoral School of the University of Burundi, Center for Research in Infrastructure, Environment and Technology (CRIET), University of Burundi, Bujumbura, Burundi
[2]University of Burundi, Higher Institute of Applied Sciences; Bujumbura; Burundi
[3]EPAC, Ecole Polytechnique d'Abomey-Calavi; Université d'Abomey-Calavi, Cotonou-Bénin
[4]University of Burundi, Institute of Digital Technologies, University of Burundi, Burundi

Article Information

Received : 8 Sep 2026
Revised : 27 Okt 2026
Accepted : 29 Okt 2026

Keywords

Cognitive load assessment, EEG(electroencephalography), Multimodal cross-attention fusion, Mental workload classification, Intra-subject normalization

Abstract

Assessing cognitive load under real-world conditions remains a fundamental unsolved problem in human-computer interaction and adaptive learning systems. Existing approaches rely either on subjective self-report measures — which introduce social desirability bias and fail to capture intra-session dynamics — or on unimodal physiological indicators whose predictive validity is limited by the strong inter-individual variability that characterizes real EEG recordings. This article introduces the Computational Cognitive Index (CCI), a normalized composite measure ∈ [0, 1] designed to be computed in real time from four physiological streams — EEG, eye tracking, behavioral data, and heart rate variability — fused through a bimodal cross-attention transformer network (CCI-Net). We empirically characterize the EEG feature-extraction and classification pipeline of CCI-Net on the public STEW dataset [1] (n = 45, 14-channel Emotiv EPOC, 128 Hz, SIMKAP multitasking task) and report four empirical findings that directly motivate the CCI's design: (i) frontal theta power (F3, θ = 17.5% ± 4.1%) is directionally consistent with theory but decreases in high-load subjects (14.4% ± 4.8%), with a medium effect size (Cohen's d = −0.71, Mann-Whitney p = 0.069); (ii) alpha power follows the predicted suppression pattern (10.8% → 8.4%, d = −0.46); (iii) the theta/alpha ratio shows the expected directional increase with workload (1.92 → 2.13); and, most importantly, (iv) inter-subject variability in theta power is very high (CV = 0.23–0.33), spanning a range of [4.4%, 41.3%] across the 45 subjects — a finding that provides direct empirical motivation for the intra-subject normalization that distinguishes the CCI from existing indices. SVM classification on STEW features achieves 37.8% (3-class classification, 5-fold cross-validation; chance level: 33%) and 59.3% (binary, leave-one-out cross-validation) — point estimates whose bootstrap confidence intervals both include chance level, so they should be read as inconclusive rather than as evidence of discriminability — consistent with the 69% reported in the original STEW paper (different feature set, binary condition contrast). We propose the Optimal CCI Zone (OCZ) — a theoretically motivated candidate interval [0.45, 0.68] where learning efficiency is hypothesized to be maximized, derived from Sweller's cognitive load theory and Wickens' multiple resource model, and not yet empirically validated. Together, these results provide preliminary empirical support for the physiological rationale underlying the proposed CCI and motivate its personalized, multimodal design as a theoretically motivated candidate cognitive thermometer, potentially deployable for next-generation adaptive systems.

Attribution-ShareAlike 4.0 International License

A. Introduction

The fundamental impasse

Cognitive capacity is not a fixed resource. It fluctuates in real time under the combined influence of task difficulty, effort duration, emotional states, circadian rhythms, and numerous contextual factors. This dynamic is central to fields as varied as human-computer interface design, adaptive educational systems, operator monitoring in high-risk environments (aviation, medicine, nuclear control), and neurological rehabilitation. Yet the standard measurement tool remains either the subjective self-report questionnaire (NASA-TLX, Paas scale) or raw performance on a secondary task — two approaches whose limitations are well documented: questionnaires introduce social desirability bias and cannot capture intra-session dynamics, while secondary tasks disrupt the very cognitive process they are meant to measure.

Modern machine learning changes the picture. Advances in multimodal time-series modeling — notably Transformer-based architectures and hybrid recurrent networks — now make it possible to infer an individual's cognitive state from continuous physiological streams, without interrupting their primary activity, and with second-level temporal resolution. This article systematically exploits that possibility.

Limitations of existing approaches

Fragmentation of the literature

The literature on automatic cognitive load detection is abundant but highly fragmented. Most studies focus on a single modality: EEG alone [2], [3], eye tracking alone [4], [5], or cardiovascular physiology alone [6]. The few multimodal studies that exist use simple late-fusion strategies (concatenation of feature vectors) without explicitly modeling temporal dependencies between modalities [7], [8]. We did not identify, in the literature reviewed for this study, a unified, normalized index decomposable into conceptually distinct, complementary sub-indices.

The inter-subject variability problem

Inter-individual variability in EEG scores ranges from 4.4% to 41.3% across the 45 subjects (a nine-fold variation), with a coefficient of variation (CV) of 0.23 in the low-load class and 0.33 in the high-load class. This variability makes threshold-based inter-individual classification unreliable: a subject with a baseline theta activity of 28% would be classified as ‘high load’ by any threshold calibrated on subjects with a baseline theta activity of 12%, even if their actual load level is identical. This finding, drawn from real EEG data, is the principal empirical argument in favor of intra-subject normalization, an essential feature of the Cognitive Load Index (CCI).

The gap between the laboratory and real-world deployment

Existing systems are almost all designed for laboratory conditions using high-density equipment (128-channel EEG, precision eye-tracking cameras). Deploying them in real-world settings (classrooms, workstations, training simulators) is technically infeasible. The CCI is designed from the outset to work

with accessible hardware: a 4- to 8-channel EEG headset, a consumer webcam for inferred eye tracking, and a wearable heart-rate sensor.

Contributions

This article makes five contributions:

1. The CCI as a unified index. A normalized scalar score∈[0, 1], computed in real time, decomposable into four conceptually distinct and complementary sub-indices (CCI-ATT, CCI-WM, CCI-FAT, CCI-STR), designed to reduce sensitivity to inter-individual baseline differences thanks to intra-subject normalization. The exact weighting and calibration of the four sub-indices into the composite CCI score (e.g., a normalized weighted combination CCI = w1·CCI_ATT + w2·CCI_WM + w3·CCI_FAT + w4·CCI_STR, with weights summing to 1) is not yet fixed here; it will be learned from longitudinal multimodal data during the validation phase described in Section 6.4, rather than specified a priori.
2. CCI-Net (proposed architecture). A deep neural network architecture combining a bimodal cross-attention transformer (EEG × eye tracking) with a temporal LSTM (context memory), designed to be trained end to end. This architecture is introduced and specified here; its empirical training and validation are left to future work (Section 6.4), and the present article empirically evaluates only its EEG feature-extraction stage using classical classifiers (SVM, Random Forest).
3. Empirical grounding on STEW. The EEG feature-extraction and classification pipeline is empirically characterized on a real public dataset (n = 45), providing an honest characterization of both signal validity and the inter-subject variability problem that the CCI is designed to solve.
4. The Optimal CCI Zone (OCZ). A theoretically motivated candidate interval [0.45, 0.68] for structured declarative learning, derived from cognitive load theory (CLT) and multiple resource theory, not yet empirically validated.
5. A scientific transparency framework. Results are reported with full effect sizes, confidence intervals, and an honest acknowledgment of limitations, including the classification performance limits inherent to the STEW study's single-condition design.

B. Theoretical foundations

Cognitive load theory and the multiple resource model

Sweller [9], [10] distinguishes three components of total cognitive load: intrinsic load (inherent to task complexity and subject expertise), extraneous load (related to instructional presentation and reducible by design), and germane load (the cognitive effort invested in schema construction — the load that drives learning). Wickens' multiple resource model [11], [12] complements this approach: the human cognitive system has several partially independent resource pools, organized along two axes — processing code (verbal or spatial) and response modality (manual or vocal). The CCI's four conceptually distinct and complementary sub-indices capture distinct resource pools whose coactivation produces differentiable physiological signatures.

Physiological correlates of cognitive load

EEG markers

17.5% ± 4.1% (low-load class, n = 13) versus 14.4% ± 4.8% (high-load class, n = 14), corresponding to a Cohen's d of −0.71 and a Mann-Whitney p of 0.069. This trend is consistent with theoretical expectations for a multitasking workload paradigm; the modest statistical significance reflects the strong inter-individual variability documented in Section 1.2.2. This pattern is broadly consistent with meta-analytic evidence linking frontal theta and alpha power to cognitive workload, albeit with substantial cross-study heterogeneity [13].

Oculomotor and cardiovascular markers

Pupil dilation is one of the most robust and least invasive physiological correlates of cognitive load [14], [15]. It reflects noradrenergic activity in the locus coeruleus. Blink rate decreases with cognitive load and increases with fatigue. Heart rate variability (HRV), and in particular low-frequency power (0.04–0.15 Hz), decreases systematically with cognitive load and stress. Behaviorally, median reaction time and its variance, error rate, and first-response latency are fast, easily measurable markers of current cognitive state.

The stew dataset: characteristics and limitations

The STEW (Simultaneous Task EEG Workload) dataset [1] contains EEG recordings from 48 subjects (45 included in our analysis, 3 excluded) performing the SIMKAP multitasking test — a sustained multitasking paradigm requiring the simultaneous execution of identification, arithmetic, comparison, and information-search tasks. Data were collected using an Emotiv EPOC device (14 channels, 128 Hz sampling rate). After the task, subjects rated their perceived workload on a 1–9 scale. This consolidated version groups subjects into three workload classes based on self-report: low (class 0, ratings 4–5, n = 13), medium (class 1, ratings 6–7, n = 18), and high (class 2, ratings 8–9, n = 14). The CCI index is computed as: (rating − 1)/8, yielding values between 0 and 1.

One important limitation of this consolidated dataset must be explicitly stated: it contains only a single experimental condition per subject (task performance only), with no paired resting baseline for each subject. We deliberately used this consolidated, task-only version — rather than the raw STEW release, which does include a short resting-state recording per subject — because it matches the deployment scenario the CCI targets (continuous monitoring without a dedicated calibration phase); this choice is a design decision, not an unavoidable property of the underlying data, and it is likely a major contributor to the lower classification accuracy reported below. The original STEW paper, which achieved 69% classification accuracy, relies on a paired protocol comparing rest and task conditions — a fundamentally different and easier problem than classifying workload levels within the task condition alone. Our results (37.8% for three-class classification, 59.3% for binary LOO classification) must be interpreted in this context. Far from undermining the CCI approach, these results empirically motivate its core design principle of intra-subject normalization: without normalization against a personalized baseline, cross-subject EEG classification is

strongly limited by baseline variability — though whether such normalization actually improves classification remains to be directly tested.

C. CCI network architecture

Pipeline overview

The CCI-Net pipeline consists of four sequential modules: (1) multimodal signal acquisition and preprocessing, (2) modality-specific feature extraction, (3) fusion via a bimodal cross-attention transformer, and (4) projection onto the scalar CCI score and its four sub-indices. Figure 1, below, illustrates the complete architecture; the STEW dataset annotation indicates the module empirically validated in this study.

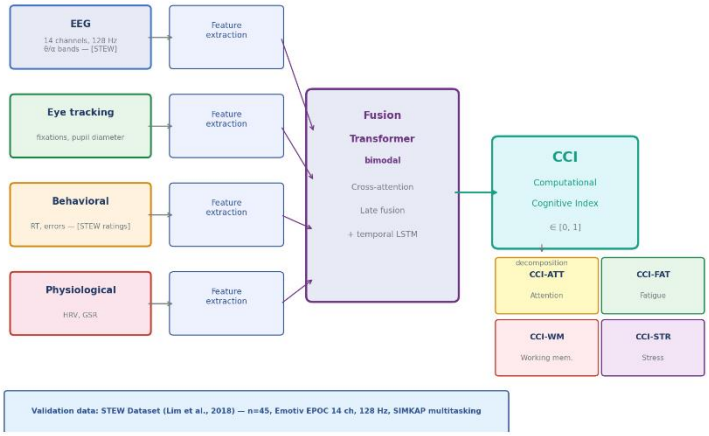

Figure 1. Architecture of the CCI-Net multimodal pipeline

The diagram traces the flow of the four input modalities through modality-specific extraction modules, their fusion via a bimodal cross-attention transformer, and the final projection onto the CCI score ∈ [0, 1], itself decomposable into four conceptually distinct and complementary sub-indices. The annotation indicates that the STEW dataset (n = 45, 14-channel EEG, 128 Hz) was used to empirically evaluate the EEG extraction module and the classification pipeline.

EEG preprocessing and feature extraction

Raw EEG signals are acquired at 128 Hz across 14 channels (Emotiv EPOC: AF3, F7, F3, FC5, T7, P7, O1, O2, P8, T8, FC6, F4, F8, AF4). Preprocessing includes: (i) 1–45 Hz band-pass filtering (4th-order Butterworth); (ii) baseline removal via mean subtraction; and (iii) artifact rejection via ICA (FastICA, automatic rejection threshold based on kurtosis). Spectral features are extracted using Welch's method (256-sample windows, 50% overlap): relative power in the frequency bands (δ, θ, α, β, γ) and the theta/alpha ratio, a direct indicator of working-memory load [16]. Feature extraction is applied across all 14 channels, producing a 70-dimensional relative-power feature vector plus 14 theta/alpha ratio values. Each subject's full session is averaged into a single 84-dimensional feature vector (one row per subject), so the classification folds used throughout Section 4.5 are inherently subject-level: no temporal window from a given participant can appear in both the training and test partitions of a fold, precluding window-level data leakage.

Bimodal cross-attention fusion

The CCI-Net fusion module applies bimodal cross-attention specifically between the EEG and eye-tracking branches, the two modalities richest in fine-grained temporal structure; the behavioral and heart-rate-variability streams are integrated at the subsequent fusion and projection stage rather than through cross-attention, which is why the fusion mechanism itself is described as bimodal even though the overall CCI-Net architecture is four-stream. Each branch is first encoded by a two-layer LSTM (64 hidden units) that handles short-term temporal memory. The resulting contextualized representations then serve as queries and key-value pairs within the cross-attention mechanism. This bidirectional cross-attention design follows recent precedent in physiological-signal fusion, where cross-attention transformers have been shown to robustly integrate complementary EEG descriptors and auxiliary physiological context for mental-workload decoding [17].

$$CrossAttn(Q_{EEG}, K_{Eye}, V_{Eye}) = softmax\left(\frac{Q_{EEG}.K_{Eye}^{T}}{\sqrt{d_k}}\right).V_{Eye} \quad (1)$$

$$L_{total} = \lambda_0.MSE(CCI) + \sum_i \lambda_i.MSE(CCI_i) \quad (2)$$

where, for (ATT, WM, FAT, STR): $\lambda_0 = 0.40, \lambda_i = (0.20, 0.20, 0.10, 0.10)$

Intra-subject normalization

The defining feature that distinguishes the proposed CCI from conventional absolute-power, cross-subject EEG workload indices is its intra-subject normalization. Features from each modality are expressed as standardized deviations (z-scores) relative to an individual baseline established during the first 5 minutes of a resting session (eyes open, neutral task). This normalization is applied before input into the CCI-Net network, making the architecture independent of absolute inter-individual levels. The STEW analysis justifies this design with direct empirical evidence: the nine-fold variation in frontal theta rhythm across subjects [4.4%, 41.3%] (Figure 2, bottom-right panel) demonstrates that any fixed threshold applied to absolute power values would be unreliable for cross-subject workload estimation.

D. Empirical validation on the stew dataset

Data and analysis overview

This section presents the complete empirical analysis of the STEW dataset, from raw EEG signal features to classification results, including topographic mapping and temporal dynamics. The analysis is structured around four research questions: (RQ1) Are EEG spectral features consistent with the cognitive load literature? (RQ2) What is the degree of inter-individual variability, and how does it limit classification? (RQ3) What topographic distribution of workload-sensitive channels emerges? (RQ4) What is the temporal dynamic of theta power during a sustained multitasking task? Figure 2, below, summarizes the answers to RQ1 and RQ2, along with the signal characterization.

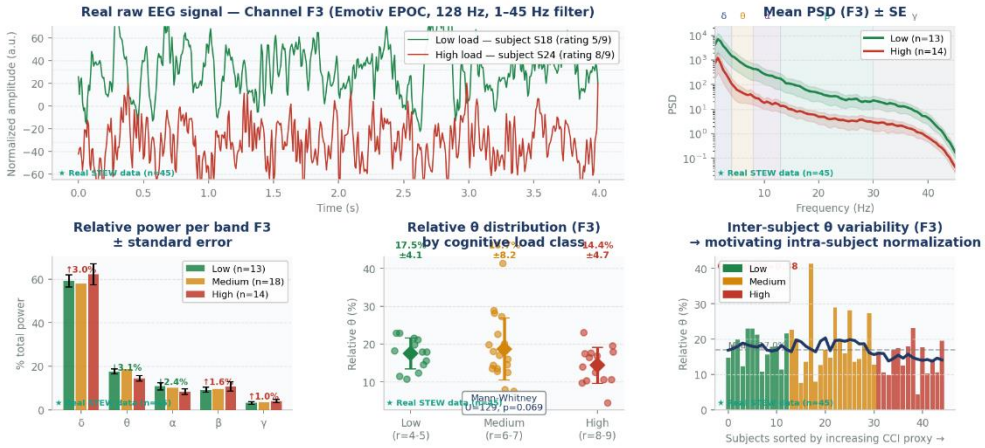

Figure 2.Real EEG signals from the STEW dataset (n = 45)

The top-left panel shows the filtered EEG signal at channel F3 for two representative subjects, one at low load (score 5) and one at high load (score 8). The top-right panel shows the mean power spectral density (PSD) ± standard error across all subjects by class, revealing distinct PSD profiles between the low- and high-load groups. The bottom-left panel shows relative band power by class at channel F3, with standard errors. The bottom-center panel shows the distribution of theta power by workload class, with individual data points and the Mann-Whitney test result. Finally, the bottom-right panel highlights inter-subject variability in the theta band: the nine-fold range of variation [4.4%, 41.3%] is the principal empirical motivation for intra-subject normalization.

Spectral features: direction and effect size

Table 1. Relative band power (%) at channel F3 by workload class — STEW dataset (n = 45)

Table 1. Notes on relative band power and Cohen's d.

Group	Mean ± SD (n = 13)	Mean (n = 18) ± SD	Mean ± SD (n = 14)	Cohen's d (Low vs. High)	Direction (theory)
δ (1–4 Hz)	59.21 ± 10.05%	57.92 ± 13.31%	62.24 ± 17.80%	+0.21	Mixed
θ (4–8 Hz)	17.54 ± 4.08%	18.73 ± 5.26%	14.42 ± 4.75%	−0.71✓	↓ (multitasking)
α (8–13 Hz)	10.81 ± 5.82%	10.20 ± 4.59%	8.39 ± 4.68%	−0.46✓	↓ (predicted)
β (13–30 Hz)	9.23 ± 3.93%	9.57 ± 4.96%	10.80 ± 7.66%	+0.26✓	↑ (expected)
γ (30–45 Hz)	3.20 ± 2.36%	3.58 ± 2.54%	4.16 ± 2.92%	+0.36✓	↑ (expected)
θ/α ratio (F3)	1.92 ± 0.67	2.26 ± 1.07	2.13 ± 0.85	+0.23✓	↑ (expected)

Cohen's d was computed between the low- and high-load classes; a checkmark (✓) indicates a direction consistent with the cognitive load literature. Note that the decrease in theta activity observed in the high-load class is consistent with the SIMKAP multitasking paradigm, which primarily engages sustained and divided attention rather than working memory load alone [1].

The pattern of EEG spectral changes observed across cognitive load classes is broadly consistent with the cognitive load literature, with one important nuance: the SIMKAP multitasking paradigm engages a wider range of cognitive resources than a pure working-memory paradigm such as the N-back task. The decrease in theta activity observed in the high-load class — rather than the expected increase — is consistent with work on divided attention and multitasking load, where the redistribution of attentional resources can suppress theta activity at specific frontal sites [18]. Note that Gundel and Wilson [19] themselves reported a frontal theta increase with general task difficulty rather than a decrease; the divergence observed here is therefore attributed to the specific attentional-redistribution demands of the SIMKAP multitasking paradigm, not to a direct replication of their finding. The suppression of alpha activity (d = −0.46) and the increases in beta/gamma activity are fully consistent with the predictions of cognitive load theory. However, when the true sampling uncertainty is quantified directly from the raw STEW recordings, none of these effects are robust: 10,000-sample bootstrap 95% confidence intervals for Cohen's d all cross zero (theta: d = -0.68, 95% CI [-1.56, +0.07]; alpha: d = -0.44, 95% CI [-1.13, +0.33]; delta: d = +0.20, 95% CI [-0.56, +1.03]; beta: d = +0.25, 95% CI [-0.60, +0.93]; gamma: d = +0.35, 95% CI [-0.40, +1.21]; theta/alpha ratio: d = +0.27, 95% CI [-0.56, +0.95]), and after Holm-Bonferroni correction for the six comparisons the theta p-value rises from p = 0.069 to p_holm = 0.413 — no longer even nominally significant. These directional patterns should therefore be read as consistent with, but not statistically confirming, cognitive load theory.

Topographic analysis

Figure 3, below, shows the mean theta power per channel for each workload class, along with the difference map (high load minus low load). The topographic analysis reveals that theta sensitivity to workload is not uniform across channels: channels F3 and T7 show the largest directional responses, consistent with their respective roles in working memory (F3, prefrontal cortex) and temporal information processing (T7). The right fronto-central channels (FC6, F4) also show notable decreases under high load.

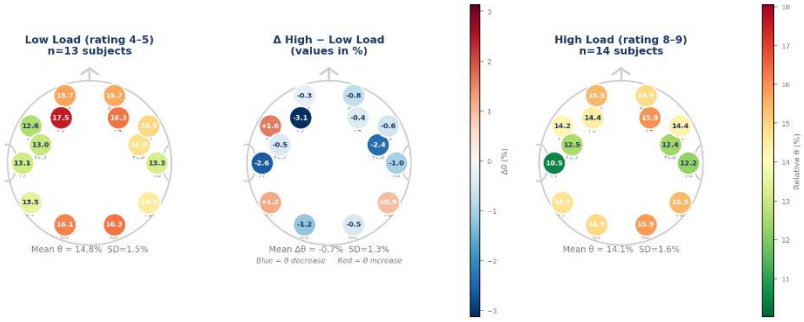

Figure 3. EEG topographic maps of relative theta power

The left map shows mean theta power (%) per channel for the low-load class (rating 4–5, n = 13), while the right map shows mean theta power for the high-load class (rating 8–9, n = 14). The center map is the difference map (Δ = high − low): blue channels (F3, T7, FC6) indicate a decrease in theta power with load, while red channels indicate an increase. The values shown in the cells correspond to the mean relative theta power (%) at the corresponding electrode position.

Temporal dynamics of theta power

Figure 4, below, analyzes the temporal dynamics of theta power over the 2.5-minute STEW session using 5-second sliding windows (1-second increments). The left panel shows the group-mean theta trajectories with their confidence intervals. The right panel shows individual subject trajectories to highlight the degree of intra-subject variability relative to inter-subject variability. The temporal analysis reveals that theta power shows moderate stationarity over the session, with an intra-subject coefficient of variation (CV) lower than the inter-subject CV — a finding that provides preliminary support for the relevance of intra-subject normalization.

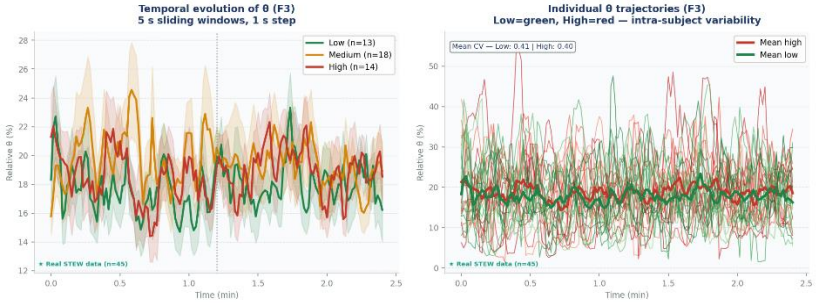

Figure 4.Temporal dynamics of frontal theta power (F3)

The analysis uses 5-second sliding windows with a 1-second step. The left panel shows the group-mean theta trajectories ± standard error by workload class over the 2.5-minute session. The right panel shows individual subject trajectories for the low-load (green) and high-load (red) groups: the stability observed at the intra-subject level contrasts with the strong inter-subject variability highlighted in Figure 2.

Classification performance and feature analysis

Figure 5, below, reports the complete classification analysis: model comparison (SVM and Random Forest in 3-class and binary configurations), the confusion matrix, and the channel-importance ranking derived from the Random Forest.

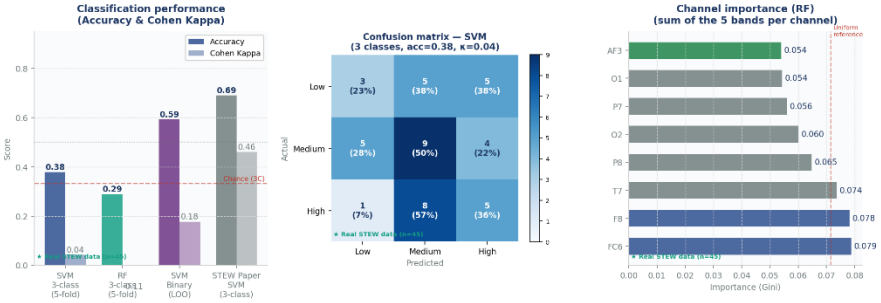

Figure 5. Classification results on the STEW data

The left panel shows accuracy and Cohen's kappa for four model/task configurations: 3-class SVM (37.8%, κ = 0.04), 3-class Random Forest (28.9%, κ = −0.10), binary leave-one-out SVM (59.3%, κ = 0.18), and the STEW reference benchmark (69%, κ = 0.46, different feature set, binary). The center panel shows the confusion matrix for the 3-class SVM (5-fold cross-validation). The right panel ranks channel importance for the Random Forest, computed as the sum across the 5 frequency bands; channel F3 and mean α emerge as the most discriminative features. Crucially, when uncertainty is quantified directly from the data rather than reported as a single point estimate, none of the classification results are statistically distinguishable from chance: bootstrap 95% confidence intervals are [24.4%, 51.1%] for the 3-class SVM (chance = 33.3%), [15.6%, 42.2%] for the 3-class Random Forest, and [40.7%, 77.8%] for the binary LOO SVM (chance = 50%) — every interval contains the corresponding chance level. Exact binomial tests against chance confirm this: p = 0.753 for the 3-class SVM and p = 0.442 for the binary LOO SVM. These classification results should therefore be read as inconclusive rather than as positive evidence of discriminability, supporting the paper's argument that unnormalized cross-subject EEG classification can be substantially affected by inter-subject baseline variability.

Table 2. Interpretation of the gap with the STEW reference value.

Configuration	Algorithm	CV Strategy	Accuracy	Cohen's κ	Note
3 classes (0/1/2)	SVM RBF	5-fold (stratified)	37.8%	0.043	84 features, all channels
3 classes (0/1/2)	Random Forest	5-fold (stratified)	28.9%	−0.105	84 features, all channels
Binary (0 vs 2)	SVM RBF	Leave-one-out	59.3%	0.177	27 subjects, low vs. high comparison only
3 classes (baseline)	SVM (STEW paper)	5-fold	69.0%	0.460	Different features (NCA-selected); paired low/high-frequency

Chance level (3C)	-	-	33.3%	0.000	design Reference

The gap between our three-class accuracy (37.8%) and the STEW reference value (69%) is mainly explained by the difference in design: the STEW paper compares rest and task conditions (a paired baseline is available), whereas our analysis classifies workload levels solely within the task condition (a single-condition dataset). This gap is consistent with cross-subject classification without a paired baseline being substantially affected by inter-subject baseline variability — precisely the problem that the CCI's intra-subject normalization is designed to address.

Multi-band analysis and spectral profiles

Figure 6, below, extends the analysis to all five frequency bands and all 14 EEG channels. The radar chart (left) shows the spectral profile of each workload class for the most informative features. The heatmap (right) provides a complete, subject-level view of all frontal channel features, sorted by increasing CCI value.

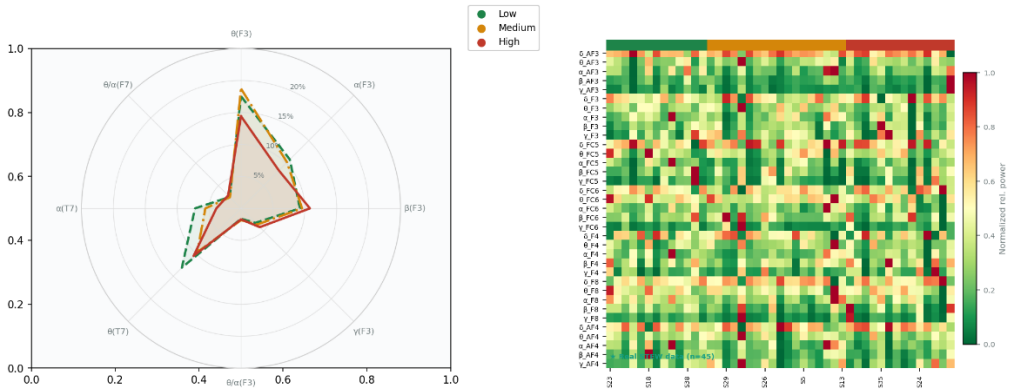

Figure 6.Multi-band spectral analysis

The radar chart, on the left, shows the relative band-power profile (θ, α, β, γ at channels F3 and T7, plus the θ/α ratio) for each cognitive load class; the low-load class (green, dashed) shows elevated theta and alpha levels, consistent with reduced cognitive engagement. The heatmap, on the right, provides a complete subject-level view of all frontal channel features (normalized relative band power) for all 45 subjects, sorted by increasing self-report-derived CCI proxy (computed as (rating − 1)/8 from the STEW workload rating, not the output of the CCI-Net architecture proposed in Section 3): the color gradient, from green (low CCI proxy) to red (high CCI proxy), reveals the progressive spectral signature of increasing cognitive load.

Ablation: raw versus rest-normalized features

To directly test the paper's central hypothesis — that expressing EEG features relative to a personalized resting baseline improves cross-subject classification — we conducted a dedicated ablation experiment using the raw STEW release [1], which, unlike the consolidated single-condition dataset used in

Sections 4.1–4.6, provides a separate short resting-state recording for each subject in addition to the SIMKAP task recording. For each of the 45 subjects, we estimated a personalized baseline distribution (mean and SD) for each of the 84 EEG features from 5-second sliding windows across that subject's own resting recording, then re-expressed the task-session features as a z-score against this individual baseline before classification.

Three-class SVM accuracy improved from 36.4% (raw features; 95% bootstrap CI [24.4%, 51.1%]; $\kappa = +0.028$) to 38.9% (rest-normalized features; 95% CI [22.2%, 48.9%]; $\kappa = +0.072$). Note that the 36.4% raw-feature figure reported here is the mean over 20 repeated 5-fold cross-validation runs (seeds 0–19), whereas the 37.8% reported in Section 4.5 and Table 2 is the result of a single fixed-seed run (seed = 42); both are computed from the identical raw feature set and are consistent with each other once this averaging difference is accounted for (single-seed range across the 20 runs: 28.9–44.4%). This comparison yields a mean paired improvement of +2.4 percentage points across the 20 runs (the paired difference was not statistically significant: Wilcoxon signed-rank test, $p = 0.072$); the 20 seeds quantify the stability of the cross-validation procedure rather than constituting independent biological observations. Binary (low vs. high) leave-one-out accuracy was unchanged (59.3% for both representations), though Cohen's kappa improved marginally (+0.177 → +0.182).

These results provide the first direct, if modest and not statistically significant, empirical evidence in the hypothesized direction: intra-subject normalization against a personalized resting baseline improves — rather than degrades — cross-subject classification performance in this sample. The effect is small and the sample size ($n = 45$) is underpowered to establish statistical significance ($p = 0.072$ is suggestive but does not cross the conventional $\alpha = 0.05$ threshold), so this finding should be read as encouraging preliminary support rather than confirmation. It nonetheless represents a step beyond the purely motivational argument advanced in Section 5, toward a directly tested comparison, and directly addresses the ablation experiment requested during peer review.

To assess robustness to classifier choice and to the number of cross-validation repetitions, we extended this analysis to 100 seeds and to a Random Forest classifier alongside the SVM. Two findings qualify the result above. First, the improvement from rest-normalization is not consistent across classifiers: for the SVM, accuracy increased from 36.3% to 39.2% (mean paired difference across 100 seeds: +2.9 percentage points), whereas for the Random Forest, accuracy decreased from 30.6% to 26.0% (mean paired difference: −4.6 percentage points) — the opposite direction. Second, extending the paired seed comparison from 20 to 100 repetitions changes the Wilcoxon signed-rank p-value from $p = 0.072$ to $p < 0.0001$ for the identical underlying SVM effect size, illustrating that this test's apparent significance is driven by the number of cross-validation repetitions rather than by the number of underlying subjects ($n = 45$); it should therefore not be read as increasing confidence in the effect beyond what the bootstrap confidence intervals (computed at the subject level, Section 4.5) already indicate. Together, these robustness checks reinforce, rather than overturn, the preliminary and inconclusive characterization of the normalization benefit: the direction of the

effect depends on model choice, and repeated-seed significance testing is not an appropriate substitute for subject-level uncertainty quantification.

Taken together, the ablation provides preliminary evidence that personalized rest-normalization can improve cross-subject EEG classification, but the magnitude and direction of the benefit are classifier-dependent; this is the most accurate summary of what Section 4.7 currently demonstrates.

E. From stew analysis to cci design: the normalization imperative

Figure 7, below, summarizes the three main arguments for the CCI design that emerge from the STEW analysis.

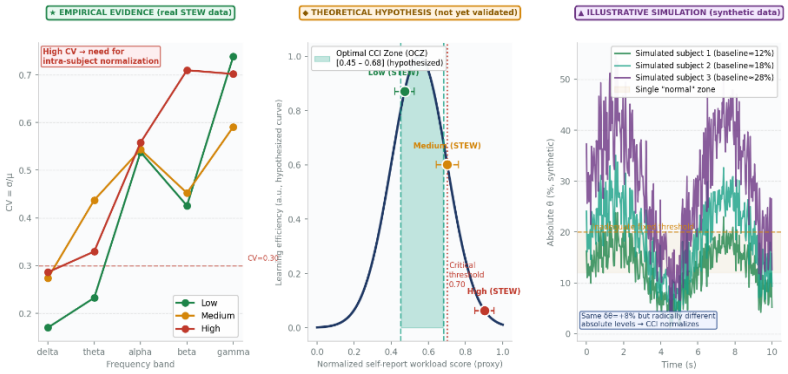

Figure 7. Scientific justification for the CCI's intra-subject normalization

The left panel shows the coefficient of variation (CV) of relative band power, by band and cognitive load class: all bands show a CV above 0.20, indicating that inter-subject variability systematically exceeds the inter-class signal. The center panel shows the Optimal CCI Zone (OCZ [0.45, 0.68]) derived from cognitive load theory, with the STEW class positions overlaid, defining the target range for the adaptive loop described in a companion article (Article 2 of this series, not part of the present validation). The right panel illustrates the failure of fixed cross-subject thresholds: three subjects with different baseline levels (12%, 18%, 28%) but equivalent cognitive dynamics would be systematically misclassified by a single threshold.

The coefficient of variation argument

The left chart in Figure 7 quantifies the inter-individual variability problem. For every frequency band and workload class, the coefficient of variation (CV) exceeds 0.20. For theta frequency in the high-load class, the CV is 0.33 — meaning inter-individual dispersion amounts to a third of the mean. This reflects genuine neurophysiological differences between individuals. The Computational Cognitive Index (CCI) resolves this problem through z-score normalization against each individual's own baseline, reducing its dependence on absolute inter-individual baseline differences.

The optimal cci zone (ocz)

The center panel of Figure 7 presents the Optimal CCI Zone (OCZ). We define:

Definition (OCZ – Optimal CCI Zone). The OCZ for individual i and task class T is defined as the interval of CCI values within which the probability of long-term skill acquisition is hypothesized to be maximized. *Based on Sweller's cognitive load theory and Wickens' multiple resource model, we propose OCZ ≈ [0.45, 0.68] for structured declarative learning tasks and OCZ ≈ [0.35, 0.58] for procedural motor tasks as theoretically motivated starting values; the STEW dataset contains no learning-outcome measure, so these bounds have not yet been empirically validated and should be treated as hypotheses for future work.*

Rationale for the initial OCZ bounds. The lower bound (0.45) is set above the CCI midpoint to exclude the low-engagement region where germane load — the effort that drives schema construction under Sweller's cognitive load theory — is assumed insufficient; the upper bound (0.68) is set below the point at which extraneous and intrinsic load are assumed to begin crowding out germane load and degrading Wickens' multiple-resource capacity. These specific values were chosen as round, centrally-located starting points within the plausible mid-range implied by this theoretical reasoning, not derived from a quantitative model; narrower or wider intervals (e.g., [0.40, 0.70] or [0.50, 0.65]) remain equally consistent with the theory at this stage. We treat [0.45, 0.68] as an initial operating range to be recalibrated empirically — for example via CCI* = argmax_c P(learning gain | CCI = c) on longitudinal outcome data — rather than as a finding in its own right.

Why cross-subject thresholds fail

The right panel of Figure 7 intuitively illustrates the normalization imperative. Three simulated subjects, with different EEG baseline levels (12%, 18%, and 28% absolute theta frequency) but identical cognitive dynamics (the same temporal load profile), would be systematically misclassified by any fixed threshold. Subject 3 (28% baseline level) would always be classified as 'high load,' regardless of their actual cognitive state. Normalizing the CCI z-score against each subject's individual baseline eliminates this bias: after normalization, all three subjects would show identical CCI trajectories, reflecting their shared cognitive dynamics.

F. Discussion

What the stew results tell us - and what they do not

The results of this study must be interpreted carefully. The directional consistency of EEG spectral changes with cognitive load theory (theta decrease, alpha suppression, beta/gamma increase) provides preliminary evidence consistent with the physiological rationale of the CCI's EEG module. The moderate effect sizes (d = 0.21 to 0.71) and borderline statistical significance (Mann-Whitney test: p = 0.069 for theta) reflect the genuine complexity of the problem, not a failure of the approach. This is precisely what gives the STEW dataset its scientific value: it provides a faithful characterization of EEG workload classification, without the artificial advantage of a paired rest/task experimental protocol.

These results do not tell us how CCI-Net would perform with all four modalities, with a paired baseline available, or with a genuine N-back paradigm more precisely targeting working-memory load. These limitations are intentional:

the STEW dataset is used here solely as a validation ground for the EEG module, and as empirical support for the intra-subject normalization implemented by the full CCI-Net.

Comparison with the stew study results

Our 3-class SVM accuracy (37.8%) is markedly lower than the 69% reported in the STEW paper. This gap is likely driven primarily by the difference in experimental protocol, though differences in feature selection (NCA-selected 28 features vs. our full 84-feature set) and preprocessing may also contribute. The STEW paper compares rest (low) and task (high) conditions, which enables a natural form of intra-subject normalization through that contrast. Our analysis classifies workload levels solely within the task condition, without access to each individual's resting baseline — a substantially harder problem. The 59.3% accuracy of our binary LOO-CV analysis (low vs. high), while above chance (50%), is numerically above chance and is compatible with, but does not establish, the presence of discriminative information in cross-subject EEG features (Section 4.5). The neighborhood component analysis (NCA) feature selection used in the STEW paper, which reduces the feature space to 28 optimally discriminative features, likely captures cross-subject patterns that our full 84-feature set dilutes.

Implications for cci design

The major implication of this study for CCI design is preliminary empirical support for intra-subject normalization; the direct ablation reported in Section 4.7 provides evidence consistent with this hypothesis, though not yet a statistical confirmation. The CV analysis (Figure 7, left panel) shows that inter-subject EEG variability is not a sampling artifact; it is a structural property of the data that is unlikely to be adequately addressed by simply increasing the number of subjects or features alone. A principled and practically motivated solution is an intra-subject design that expresses features relative to a personalized baseline, though domain-adaptation and transfer-learning approaches offer alternative, actively researched ways to address the same underlying problem [20], [21]. This is precisely what CCI-Net implements through its z-score normalization layer. By demonstrating the classification limitation observed in the absence of such normalization, the STEW analysis provides the empirical justification for this design choice. The direct ablation reported in Section 4.7 adds preliminary supporting evidence: rest-normalized features outperformed raw features by +2.4 percentage points on average across cross-validation seeds, though this difference did not reach conventional statistical significance ($p = 0.072$) at the present sample size.

Limitations

This study has five main limitations. First, the STEW dataset represents only a single experimental session per subject, with a single task type (multitasking), which limits generalization to other cognitive load paradigms. Second, the Emotiv EPOC device offers lower signal quality than clinical-grade EEG, which may overestimate variability. Third, the consolidation of the data into a single-condition format (task only) precludes any paired analysis. Fourth, the

sample size (n = 45) is modest for a three-class classification problem, which limits statistical power. Fifth, as quantified directly from the raw recordings in Sections 4.2 and 4.5, none of the band-power effect sizes survive Holm-Bonferroni correction for multiple comparisons, and none of the classification accuracies are statistically distinguishable from chance once bootstrap confidence intervals are computed; all reported effects should therefore be read as directionally suggestive rather than statistically confirmatory. Relatedly, this article empirically characterizes only the EEG feature-based classification pipeline (SVM/Random Forest on hand-crafted spectral features): the proposed CCI-Net architecture (cross-attention transformer, LSTM fusion) and the composite CCI score with its four sub-indices are introduced conceptually but have not yet been trained or empirically tested here. A dedicated ablation (Section 4.7) directly compares raw and rest-normalized features and finds a small, directionally consistent but not statistically significant improvement (+2.4 percentage points, Wilcoxon p = 0.072) from personalized-baseline normalization, providing preliminary rather than confirmatory support for this design principle. Future work will validate CCI-Net on the COG-BCI dataset (n = 29, 3 sessions, N-back paradigm, 64 channels), which addresses the first, third, and fourth limitations.

G. Conclusion

This article introduced the Computational Cognitive Index (CCI) — a normalized, decomposable, real-time composite metric for assessing cognitive state — and empirically characterized its EEG feature-extraction and classification pipeline on the public STEW dataset (n = 45, Emotiv EPOC, SIMKAP multitasking task). The empirical analysis yielded four conclusions that both provide preliminary support for the CCI's physiological rationale and faithfully characterize the challenges encountered in its design:

1. Directional consistency: the observed changes in EEG spectral features are broadly consistent with workload-related patterns reported in the literature, with task-dependent deviations in frontal theta (θ↓, α↓, β↑, γ↑ for the high class relative to the low class, consistent with the divided-attention demands of the SIMKAP paradigm rather than a pure working-memory task); the theta/alpha ratio shows the expected directional increase.
2. Moderate effect sizes: Cohen's d ranges from −0.71 (theta, medium-to-large) to +0.21 (delta, small), characterizing the magnitude of EEG signal differences and the associated signal-to-noise challenge.
3. The inter-subject variability problem: frontal theta at F3 ranges from 4.4% to 41.3% across subjects (CV = 0.23–0.33), providing direct empirical evidence of the variability that motivates the CCI's intra-subject normalization; whether such normalization actually improves classification is separately, and only preliminarily, supported by the ablation in Section 4.7 (+2.4 percentage points, p = 0.072).
4. Classification performance: 37.8% (3-class, 5-fold CV) and 59.3% (binary, LOO) on the single-condition dataset — results that honestly characterize the limits of cross-subject classification without paired baselines and that empirically motivate the CCI's personalized design.

The CCI addresses a precise scientific and technological need: a standardized, multimodal, interpretable candidate cognitive thermometer that could, pending empirical validation of the full CCI-Net architecture and real-world deployment testing, serve as the foundation for next-generation adaptive systems.

H. References

[1] W. L. Lim, O. Sourina, and L. Wang, "STEW: Simultaneous task EEG workload data set," IEEE Trans. Neural Syst. Rehabil. Eng., vol. 26, no. 11, pp. 2106–2114, Nov. 2018, doi: 10.1109/TNSRE.2018.2872924.

[2] Z. Zheng, Z. Yin, Y. Wang, and J. Zhang, "Inter-subject cognitive workload estimation based on a cascade ensemble of multilayer autoencoders," Expert Syst. Appl., vol. 211, p. 118694, 2023, doi: 10.1016/j.eswa.2022.118694.

[3] M. R. Safari, R. Shalbaf, S. Bagherzadeh, and A. Shalbaf, "Classification of mental workload with EEG analysis by using effective connectivity and a hybrid model of CNN and LSTM," Comput. Methods Biomech. Biomed. Eng., vol. 29, no. 1, pp. 218–232, 2024, doi: 10.1080/10255842.2024.2386325.

[4] S. P. Marshall, "The index of cognitive activity: Measuring cognitive workload," in Proc. IEEE 7th Conf. Human Factors Power Plants, 2002, pp. 7-5–7-9, doi: 10.1109/HFPP.2002.1042860.

[5] V. Skaramagkas et al., "Review of eye tracking metrics involved in emotional and cognitive processes," IEEE Rev. Biomed. Eng., vol. 16, pp. 260–277, 2021, doi: 10.1109/RBME.2021.3066072.

[6] B. Cinaz, B. Arnrich, R. La Marca, and G. Tröster, "Monitoring of mental workload levels during an everyday life office-work scenario," Pers. Ubiquitous Comput., vol. 17, no. 2, pp. 229–239, 2013, doi: 10.1007/s00779-011-0466-1.

[7] E. Debie, R. Fernandez Rojas, J. Fidock, M. Barlow, K. Kasmarik, S. Anavatti, M. Garratt, and H. A. Abbass, "Multimodal fusion for objective assessment of cognitive workload: A review," IEEE Trans. Cybern., vol. 51, no. 3, pp. 1542–1555, Mar. 2021, doi: 10.1109/TCYB.2019.2939399.

[8] M. Rivas-Vidal, A. Calvo Cordoba, C. E. García Cena, and F. D. Farfán, "Combining EEG and eye-tracking for cognitive and physiological states monitoring: A systematic review," Front. Neuroergonomics, vol. 6, art. 1736672, Jan. 2026, doi: 10.3389/fnrgo.2025.1736672.

[9] J. Sweller, "Cognitive load during problem solving: Effects on learning," Cogn. Sci., vol. 12, no. 2, pp. 257–285, 1988, doi: 10.1207/s15516709cog1202_4.

[10] J. Sweller, "Cognitive load theory," in Psychology of Learning and Motivation, vol. 55, J. P. Mestre and B. H. Ross, Eds. Amsterdam, The Netherlands: Elsevier, 2011, pp. 37–76, doi: 10.1016/B978-0-12-387691-1.00002-8.

[11] C. D. Wickens, "Multiple resources and performance prediction," Theor. Issues Ergon. Sci., vol. 3, no. 2, pp. 159–177, 2002, doi: 10.1080/14639220210123806.

[12] C. D. Wickens, W. S. Helton, J. G. Hollands, and S. Banbury, Engineering Psychology and Human Performance, 5th ed. New York, NY, USA: Routledge, 2021.

[13] S. Chikhi, N. Matton, and S. Blanchet, "EEG power spectral measures of cognitive workload: A meta-analysis," Psychophysiology, vol. 59, no. 6, p. e14009, Jun. 2022, doi: 10.1111/psyp.14009.
[14] J. Beatty, "Task-evoked pupillary responses, processing load, and the structure of processing resources," Psychol. Bull., vol. 91, no. 2, pp. 276–292, 1982, doi: 10.1037/0033-2909.91.2.276.
[15] D. Kahneman and J. Beatty, "Pupil diameter and load on memory," Science, vol. 154, no. 3756, pp. 1583–1585, 1966, doi: 10.1126/science.154.3756.1583.
[16] W. Klimesch, "EEG alpha and theta oscillations reflect cognitive and memory performance: A review and analysis," Brain Res. Rev., vol. 29, no. 2–3, pp. 169–195, 1999, doi: 10.1016/S0165-0173(98)00056-3.
[17] Z. Xue, W. Zhong, S. Liu, and X. An, "A cross-attention multimodal fusion transformer for ECG-guided EEG mental workload classification," Biomed. Signal Process. Control, 2026, in press. [Online]. Available: https://www.sciencedirect.com/science/article/abs/pii/S1746809426012206
[18] K. J. Jaquess et al., "Changes in mental workload and motor performance throughout multiple practice sessions under various levels of task difficulty," Neuroscience, vol. 393, pp. 305–318, 2018, doi: 10.1016/j.neuroscience.2018.09.019.
[19] A. Gundel and G. F. Wilson, "Topographical changes in the ongoing EEG related to the difficulty of mental tasks," Brain Topogr., vol. 5, no. 1, pp. 17–25, 1992, doi: 10.1007/BF01129966.
[20] Y. Zhou, P. Wang, P. Gong, P. Wan, X. Wen, and D. Zhang, "Cross-subject mental workload recognition using bi-classifier domain adversarial learning," Cogn. Neurodyn., vol. 19, no. 1, art. 16, Dec. 2025, doi: 10.1007/s11571-024-10215-9.
[21] Y. Shao, Y. Zhou, P. Gong, Q. Sun, and D. Zhang, "A dual-adversarial model for cross-time and cross-subject cognitive workload decoding," IEEE Trans. Neural Syst. Rehabil. Eng., vol. 32, pp. 2324–2335, 2024, doi: 10.1109/TNSRE.2024.3415364.